METHOD FOR SCANNING, ANALYZING AND HANDLING VARIOUS KINDS OF DIGITAL INFORMATION CONTENT

RELATED APPLICATION DATA

[0001] This application is a continuation of Serial No. 60/060,610 filed 10/1/97 and incorporated herein by this reference.

TECHNICAL FIELD

[0002] The present invention pertains to methods for scanning and analyzing various kinds of digital information content, including information contained in web pages, email and other types of digital datasets, including multi-media datasets, for detecting specific types of content. As one example, the present invention can be embodied in software for use in conjunction with web browsing software to enable parents and guardians to exercise control over what web pages can be downloaded and viewed by their children.

BACKGROUND OF THE INVENTION

[0003] Users of the World-Wide Web ("Web") have discovered the benefits of simple, low-cost global access to a vast and exponentially growing repository of information, on a huge range of topics. Though the Web is also a delivery medium for interactive computerized applications (such as online airline travel booking systems), a major part of its function is the delivery of information in response to a user's inquiries and ad-hoc exploration - a process known popularly as "surfing the Web."

[0004] The content delivered via the Web is logically and semantically organized as "pages" - autonomous collections of data delivered as a package upon request.

Web pages typically use the HTML language as a core syntax, though other delivery syntaxes are available.

[0005] Web pages consist of a regular structure, delineated by alphanumeric commands in HTML, plus potentially included media elements (pictures, movies, sound files, Java programs, etc.). Media elements are usually technically difficult or time-consuming to analyze.

[0006] Pages were originally grouped and structured on Web sites for publication; recently, other forms of digital data, such as computer system file directors, have also been made accessible to Web browsing software on both a local and shared basis.

[0007] Another discrete organization of information which is analogous to the Web page is an individual email document. The present invention can be applied to analyzing email content as explained later.

[0008] The participants in the Web delivery system can be categorized as publishers, who use server software and hardware systems to provide interactive Web pages, and end-users, who use web-browsing client software to access this information. The Internet, tying together computer systems worldwide via interconnected international data networks, enables a global population of the latter to access information made available by the former. In the case of information stored on a local computer system, the publisher and end-user may clearly be the same person but given shared use of computing resources, this is not always so.

[0009] The technologies originally developed for the Web are also being increasingly applied to the local context of the personal computer environment, with Web-browsing software capable of viewing and operating on local files. This patent application is primarily focused on the Web-based environment, but also envisions the applicability

of many of the petitioners' techniques to information bound to the desktop context.

[0010] End-users of the Web can easily access many dozens of pages during a single session. Following links from search engines, or from serendipitous clicking of the Web links typically bound within Web pages by their authors, users cannot anticipate what information they will next be seeing.

[0011] The data encountered by end-users surfing the Web takes many forms. Many parents are concerned about the risk of their children encountering pornographic material online. Such material is widespread. Other forms of content available over the Web create similar concern, including racist material and hate-mongering, information about terrorism and terrorist techniques, promotion of illicit drugs, and so forth. Some users may not be concerned about protecting their children, but rather simply wish themselves not to be inadvertently exposed to offensive content. Other persons have managerial or custodial responsibility for the material accessed or retrieved by others, such as employees; liability concerns often arise from such access.

SUMMARY OF THE INVENTION

[0012] In view of the foregoing background, one object of the present invention is to enable parents or guardians to exercise some control over the web page content displayed to their children.

[0013] Another object of the invention is to provide for automatic screening of web pages or other digital content.

[0014] A further object of the invention is to provide for automatic blocking of web pages that likely include pornographic or other offensive content.

[0015] A more general object of the invention is to characterize a specific category of information content by example, and then to efficiently and accurately identify instances of that category within a real-time data stream.

[0016] A further object of the invention is to support filtering, classifying, tracking and other applications based on real-time identification of instances of particular selected categories or content - with or without displaying that content.

[0017] The invention is useful for a variety of applications, including but not limited to blocking digital content, especially world-wide web pages, from being displayed when the content is unsuitable or potentially harmful to the user, or for any other reason that one might want to identify particular web pages based on their content.

[0018] According to one aspect of the invention, a method for controlling access to potentially offensive or harmful web pages includes the following steps: First, in conjunction with a web browser client program executing on a digital computer, examining a downloaded web page before the web page is displayed to the user. This examining step includes identifying and analyzing the web page natural language content relative to a predetermined database of words - or more broadly regular expressions - to form a rating. The database or "weighting list" includes a list of expressions previously associated with potentially offensive or harmful web pages, for example pornographic pages, and the database includes a relative weighting assigned to each word in the list for use in forming the rating.

[0019] The next step is comparing the rating of the downloaded web page to a predetermined threshold rating. The threshold rating can be by default, or can be selected,

for example based on the age or maturity of the user, or other "categorization" of the user, as indicated by a parent or other administrator. If the rating indicates that the downloaded web page is more likely to be offensive or harmful than a web page having the threshold rating, the method calls for blocking the downloaded web page from being displayed to the user. In a presently preferred embodiment, if the downloaded web page is blocked, the method further calls for displaying an alternative web page to the user. The alternative web page can be generated or selected responsive to a predetermined categorization of the user like the threshold rating. The alternative web page displayed preferably includes an indication of the reason that the downloaded web page was blocked, and it can also include one or more links to other web pages selected as age-appropriate in view of the categorization of the user. User login and password procedures are used to establish the appropriate protection settings.

[0020] Of course the invention is fully applicable to digital records or datasets other than web pages, for example files, directories and email messages. Screening pornographic web pages is described to illustrate the invention and it reflects a commercially available embodiment of the invention.

[0021] Another aspect of the invention is a computer program. It includes first means for identifying natural language textual portions of a web page and forming a list of words or other regular expressions that appear in the web page; a database of predetermined words that are associated with the selected characteristic; second means for querying the database to determine which of the list of words has a match in the database; third means for acquiring a corresponding weight from the database for each such word having a match in the database so as to form a

weighted set of terms; and fourth means for calculating a rating for the web page responsive to the weighted set of terms, the calculating means including means for determining and taking into account a total number of natural language words that appear in the identified natural language textual portions of the web page.

[0022] As alluded to above, statistical analysis of a web page according to the invention requires a database or attribute set, compiled from words that appear in know "bad" - e.g. pornographic, hate-mongering, racist, terrorist, etc. - web pages. The appearance of such words in a downloaded page under examination does not necessarily indicate that the page is "bad," but it increases the probability that such is the case. The statistical analysis requires a "weighting" be provided for each word or phrase in a word list. The weightings are relative to some neutral value so the absolute values are unimportant. Preferably, positive weightings are assigned to words or phrases that are more likely to (or even uniquely) appear in the selected type of page such as a pornographic page, while negative weightings are assigned to words or phrases that appear in non-pornographic pages. Thus, when the weightings are summed in calculating a rating of a page, the higher the value the more likely the page meets the selected criterion. If the rating exceeds a selected threshold, the page can be blocked.

[0023] A further aspect of the invention is directed to building a database or target attribute set. Briefly, a set of "training datasets" such as web pages are analyzed to form a list of regular expressions. Pages selected as "good" (non-pornographic, for example) and pages selected as "bad" (pornographic) are analyzed, and rate of occurrence data is statistically analyzed to identify the expressions (e.g. natural language words or phrases) that

are helpful in discriminating the content to be recognized. These expressions form the target attribute set.

[0024] Then, a neural network approach is used to assigned weightings to each of the listed expressions. This process uses the experience of thousands of examples, like web pages, which are manually designated simply as "yes" or "no" as further explained later.

[0025] Additional objects and advantages of this invention will be apparent from the following detailed description of preferred embodiments thereof which proceeds with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] FIG. 1 is a flow diagram illustrating operation of a process according to the present invention for blocking display of a web page or other digital dataset that contains a particular type of content such as pornography.

[0027] FIG. 2 is a simplified block diagram of a modified neural network architecture for creating a weighted list of regular expressions useful in analyzing content of a digital dataset.

[0028] FIG. 3 is a simplified diagram illustrating a process for forming a target attribute set having terms that are indicative of a particular type of content, based on a group of training datasets.

[0029] FIG. 4 is a flow diagram illustrating a neural network based adaptive training process for developing a weighted list of terms useful for analyzing content of web pages or other digital datasets.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT

[0030] Figure 1 is a flow diagram illustrating operation of a process for blocking display of a web page

(or other digital record) that contains a particular type of content. As will become apparent from the following description, the methods and techniques of the present invention can be applied for analyzing web pages to detect any specific type of selected content. For example, the invention could be applied to detect content about a particular religion or a particular book; it can be used to detect web pages that contain neo-Nazi propaganda; it can be used to detect web pages that contain racist content, etc. The presently preferred embodiment and the commercial embodiment of the invention are directed to detecting pornographic content of web pages. The following discussions will focus on analyzing and detecting pornographic content for the purpose of illustrating the invention.

[0031] In one embodiment, the invention is incorporated into a computer program for use in conjunction with a web browser client program for the purpose of rating web pages relative to a selected characteristic -- pornographic content, for example - and potentially blocking display of that we page on the user's computer if the content is determined pornographic. In Figure 1, the software includes a proxy server 10 that works upstream of and in cooperation with the web browser software to receive a web page and analyze it before it is displayed on the user's display screen. The proxy server thus provides an HTML page 12 as input for analysis. The first analysis step 14 calls for scanning the page to identify the regular expressions, such as natural language textual portions of the page. For each expression, the software queries a pre-existing database 30 to determine whether or not the expression appears in the database. The database 30, further described later, comprises expressions that are useful in discriminating a specific category of information

such as pornography. This query is illustrated in Figure 1 by flow path 32, and the result, indicating a match or no match, is shown at path 34. The result is formation of a "match list" 20 containing all expressions in the page 12 that also appear in the database 30. For each expression in the match list, the software reads a corresponding weight from the database 30, step 40, and uses this information, together with the match list 20, to form a weighted list of expressions 42. This weighted list of terms is tabulated in step 44 to determine a score or rating in accordance with the following formula:

$$\text{rating} = \left(n\sum_{1}^{p}(x_p w_p)\right)/c$$

[0032] In the above formula, "n" is a modifier or scale factor which can be provided based on user history. Each term x_p w_p is one of the terms from the weighted list 42. As shown in the formula, these terms are summed together in the tabulation step 44, and the resulting sum is divided by a total word count provided via path 16 from the initial page scanning step 14. The total score or rating is provided as an output at 46.

[0033] Turning now to operation of the program from the end-user's perspective, again referring to Figure 1, the user interacts with a conventional web browser program by providing user input 50. Examples of well-known web-browser programs include Microsoft Internet Explorer and Netscape. The browser displays information through the browser display or window 52, such as a conventional PC monitor screen. When the user launches the browser program, the user logs-in for present purposes by providing a password at step 54. The user I.D. and password are used to look up applicable threshold values in step 56.

[0034] In general, threshold values are used to influence the decision of whether or not a particular digital dataset should be deemed to contain the selected category of information content. In the example at hand, threshold values are used in the determination of whether or not any particular web page should be blocked or, conversely, displayed to the user. The software can simply select a default threshold value that is thought to be reasonable for screening pornography from the average user. In a preferred embodiment, the software includes means for a parent, guardian or other administrator to set up one or more user accounts and select appropriate threshold values for each user. Typically, these will be based on the user's age, maturity, level of experience and the administrator's good judgment. The interface can be relatively simple, calling for a selection of a screening level - such as low, medium or high - or user age groups. The software can then translate these selections into corresponding rating numbers.

OPERATION

[0035] In operation, the user first logs-in with a user I.D. and password, as noted, and then interacts with the browser software in the conventional manner to "surf the web" or access any selected web site or page, for example, using a search engine or a predetermined URL. When a target page is downloaded to the user's computer, it is essentially "intercepted" by the proxy server 10, and the HTML page 12 is then analyzed as described above, to determine a rating score shown at path 46 in Figure 1. In step 60, the software then compares the downloaded page rating to the threshold values applicable to the present user. In a preferred embodiment, the higher the rating the more likely the page contains pornographic content. In

other words, a higher frequency of occurrence of "naughty" words (those with positive weights) drives the ratings score higher in a positive direction. Conversely, the presence of other terms having negative weights drives the score lower.

[0036] If the rating of the present page exceeds the applicable threshold or range of values for the current user, a control signal shown at path 62 controls a gate 64 so as to prevent the present page from being displayed at the browser display 52. Optionally, an alternative or substitute page 66 can be displayed to the user in lieu of the downloaded web page. The alternative web page can be a single, fixed page of content stored in the software. Preferably, two or more alternative web pages are available, and an age-appropriate alternative web page is selected, based on the user I.D. and threshold values. The alternative web page can explain why the downloaded web page has been blocked, and it can provide links to direct the user to web pages having more appropriate content. The control signal 62 could also be used to take any other action based on the detection of a pornographic page, such as sending notification to the administrator. The administrator can review the page and, essentially, overrule the software by adding the URL to a "do not block" list maintained by the software.

FORMULATING WEIGHTED LISTS OF WORDS AND PHRASES

[0037] Figure 2 is a simplified block diagram of a neurol-network architecture for developing lists of words and weightings according to the present invention. Here, training data 70 can be any digital record or dataset, such as database records, e-mails, HTML or other web pages, use-net postings, etc. In each of these cases, the records include at least some text, i.e., strings of ASCII

characters, that can be identified to form regular expressions, words or phrases. We illustrate the invention by describing in greater detail its application for detecting pornographic content of web pages. This description should be sufficient for one skilled in the art to apply the principles of the invention to other types of digital information.

[0038] In Figure 2, a simplified block diagram of a neurol-network shows training data 70, such as a collection of web pages. A series of words, phrases or other regular expressions is extracted from each web page and input to a neurol-network 72. Each of the terms in the list is initially assigned a weight at random, reflected in a weighted list 78. The network analyzes the content of the training data, as further explained below, using the initial weighting values. The resulting ratings are compared to the predetermined designation of each sample as "yes" or "no," i.e., pornographic or not pornographic, and error data is accumulated. The error information thus accumulated over a large set of training data, say 10,000 web pages, is then used to incrementally adjust the weightings. This process is repeated in an interactive fashion to arrive at a set of weightings that are highly predictive of the selected type of content.

[0039] Figure 3 is a flow diagram that illustrates the process for formulating weighted lists of expressions - also called target attribute set - in greater detail. Referring to Figure 3, a collection of "training pages" 82 is assembled which, again, can be any type of digital content that includes ASCII words but for illustrated is identified as a web page. The "training" process for developing a weighted list of terms requires a substantial number of samples or "training pages" in the illustrated embodiment. As the number of training pages increases, the

accuracy of the weighting data improves, but the processing time for the training process increases non-linerally. A reasonable tradeoff, therefore, must be selected, and the inventors have found in the presently preferred embodiment that the number of training pages (web pages) used for this purpose should be at least about 10 times the size of the word list. Since a typical web page contains on the order of 1,000 natural language words, a useful quantity of training pages is on the order of 10,000 web pages.

[0040] Five thousand web pages 84 should be selected as examples of "good" (i.e., not pornographic) content and another 5,000 web pages 86 selected to exemplify "bad" (i.e., pornographic) content. The next step in the process is to create, for each training page, a list of unique words and phrases (regular expressions). Data reflecting the frequency of occurrence of each such expression in the training pages is statistically analyzed 90 in order to identify those expressions that are useful for discriminating the pertinent type of content. Thus, the target attribute set is a set of attributes that are indicative of a particular type of content, as well as attributes that indicate the content is NOT of the target type. These attributes are then ranked in order of frequency of appearance in the "good" pages and the "bad" pages.

[0041] The attributes are also submitted to a Correlation Engine which searches for correlations between attributes across content sets. For example, the word "breast" appears in both content sets, but the phrases "chicken breast" and "breast cancer" appear only in the Anti-Target ("good") Content Set. Attributes that appear frequently in both sets without a mitigating correlation are discarded. The remaining attributes constitute the Target Attribute Set.

[0042] Figure 4 illustrates a process for assigning weights to the target attribute set, based on the training data discussed above. In Figure 4, the weight database 110 essentially comprises the target attribute set of expressions, together with a weight value assigned to each expression or term. Initially, to begin the adaptive training process, these weights are random values. (Techniques are known in computer science for generating random—or at least good quality, pseudo-random—numbers.) These weighting values will be adjusted as described below, and the final values are stored in the database for inclusion in a software product implementation of the invention. Updated or different weighting databases can be provided, for example via the web.

[0043] The process for developing appropriate weightings proceeds as follows. For each training page, similar to Figure 1, the page is scanned to identify regular expressions, and these are checked against the database 110 to form a match list 114. For the expressions that have a match in database 110, the corresponding weight is downloaded from the database and combined with the list of expressions to form a weighted list 120. This process is repeated so that weighted lists 120 are formed for all of the training pages 100 in a given set.

[0044] Next, a threshold value is selected—for example, low, medium or high value—corresponding to various levels of selectivity. For example, if a relatively low threshold value is used, the system will be more conservative and, consequently, will block more pages as having potentially pornographic content. This may be useful for young children, even though some non-pornographic pages may be excluded. Based upon the selected threshold level 122, each of the training pages 100 is designated as simply "good" or "bad" for training

purposes. This information is stored in the rated list 124 in Figure 4 for each of the training pages.

[0045] A neurol-network 130 receives the page rating (good or bad) via path 132 from the lists 124 and the weighted lists 120. It also accesses the weight database 110. The neurol-network then executes a series of equations for analyzing the entire set of training pages (for example, 10,000 web pages) using the set of weightings (database 110) which initially are set to random values. The network processes this data and takes into account the correct answer for each page—good or bad—from the list 124 and determines an error value. This error term is then applied to adjust the list of weights, incrementally up or down, in the direction that will improve the accuracy of the rating. This is known as a feed-forward or back-propagation technique, indicated at page 134 in the drawing. This type of neurol-network training arrangement is known in prior art for other applications. For example, a neurol-network software package called "SNNS" is available on the internet for downloading from the University of Stuttgart.

[0046] Following are a few entries from a list of regular expressions along with neural-net assigned weights:

18[\W]?years[\W]?of[\W]?age[\W]	500
adults[\W]?only[\W]	500
bestiality[\W]	250
chicken[\W]breasts? [\W]	-500
sexuality[\W]?(oriented|explicit)[\W]	500

OTHER APPLICATIONS

[0047] As mentioned above, the principles of the present invention can be applied to various applications other than web-browser client software. For example, the

present technology can be implemented as a software product for personal computers to automatically detect and act upon the content of web pages as they are viewed and automatically "file," i.e., create records comprising meta-content references to that web-page content in a user-modifiable, organizational and presentation schema.

[0048] Another application of the invention is implementation in a software product for automatically detecting and acting upon the content of computer files and directories. The software can be arranged to automatically create and record meta-content references to such files and directories in a user-modifiable, organizational and presentation schema. Thus, the technology can be applied to help end users quickly locate files and directories more effectively and efficiently than conventional directory-name and key-word searching.

[0049] Another application of the invention is e-mail client software for controlling pornographic and other potentially harmful or undesired content and e-mail. In this application, a computer program for personal computers is arranged to automatically detect and act upon e-mail content—for example, pornographic e-mails or unwanted commercial solicitations. The program can take actions as appropriate in response to the content, such as deleting the e-mail or responding to the sender with a request that the user's name be deleted from the mailing list.

[0050] The present invention can also be applied to e-mail client software for categorizing and organizing information for convenient retrieval. Thus, the system can be applied to automatically detect and act upon the content of e-mails as they are viewed and automatically file meta-content references to the content of such e-mails, preferably in a user-modifiable, organizational and presentation schema.

[0051] A further application of the invention for controlling pornographic or other undesired content appearing in UseNet news group postings and, like e-mail, the principles of the present invention can be applied to a software product for automatically detecting and acting upon the content of UseNet postings as they are received and automatically filing meta-content references to the UseNet postings in a user-modifiable, organizational and presentation schema.

[0052] It will be obvious to those having skill in the art that many changes may be made to the details of the above-described embodiment of this invention without departing from the underlying principles thereof. The scope of the present invention should, therefore, be determined only by the following claims.